to the load on SUMEX. In addition the Megatek is interfaced directly to the SUMEX 2020 rather than the more heavily loaded dual PDP-10 system. This graphics/software package, while not a major research result in itself, represents a significant improvement in the ease of use of our programs.

2. Research in Progress

The following are some highlights of research in progress. The common theme of these studies is representation of stereochemistry and use of stereochemical information in answering questions concerning the nature of known or unknown molecular structures.

2.1 Structure Design. We have frequently pointed out that while our computer programs (CONGEN, GENOA, etc.) are designed to be used for chemical structure elucidation, much of the software and code could find use in the problem of structure design. In a crude sense the problems differ only in that the design problem will generally require more than one molecular formula to start with and is expected to have more than a single structure as an "answer" or result. We have resolved the first difference with the RANGEN program which permits a range of molecular formulas (section 1.2 above). We are continuing work on the latter problem of dealing with the imposition of constraints efficiently (particularly geometric constraints) and the unmanageably large number of structures which usually result in problems of this kind. As has been our practice in the past, we are progressing with this program development while working on a real problem of biomedical significance, in this case the design of GABA (gamma-amino butyric acid) analogues. We hope to both contribute to software development and augment the list of GABA analogues by this effort.

2.2 Representation and manipulation of conformational stereochemistry. Intense efforts will continue to augment our programs to include conformational information in the structure representation. Our new programs for specification of conformational structures or substructures (section 1.3 above) will be further developed and integrated into our existing computer programs. Our efforts to generate conformations with constraints will also continue. Amongst the various alternative approaches we have chosen to proceed on this difficult problem by using the method of distance geometry. This appears to be a computationally tractable approach which can ensure an exhaustive and irredundant generation, the same standards achieved in our earlier structure generation programs (CONGEN, GENOA, STEREO). Such a computer program for conformation generation would find use both in structure elucidation and structure design.

2.3 Three-dimensional common substructures. The field of structure/activity relationships assumes that if each member of a collection of compounds exhibits a defined sort of activity (biological, spectroscopic, chemical...), there must be some common feature (substructure). The process of examining a collection of structures manually for common features is possible (particularly with sophisticated graphics terminals) but inefficient and inexact. We believe a better

approach is a computer search of the three-dimensional structures. Even this problem (an algorithm for an exhaustive search) is difficult since the problem involved is known to belong to a class of problems for which exhaustive, efficient solutions are extremely unlikely. We have implemented a preliminary method for doing this search using the methods of distance geometry (see Reference 26) and this work will continue.

D. List of Recent Publications

(1) J.G. Nourse, R.E. Carhart, D.H. Smith, and C. Djerassi, "Exhaustive Generation of Stereoisomers for Structure Elucidation," J. Am. Chem. Soc., 101, 1216 (1979).

(2) C. Djerassi, D.H. Smith, and T.H. Varkony, "A Novel Role of Computers in the Natural Products Field," Naturwiss., 66, 9 (1979).

(3) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart and B.G. Buchanan, "Use of a Computer to Identify Unknown Compounds. The Automation of Scientific Inference," Chapter 7 in "Biomedical Applications of Mass Spectrometry, First Supplementary Volume," G.R. Waller and O.C. Dermer, Eds., John Wiley and Sons, Inc., New York, 1980, p. 125.

(4) T.C. Rindfleisch, D.H. Smith, W.J. Yeager, M.W. Achenbach, and A. Wegmann, "Mass Spectrometer Data Acquisition and Processing Systems," in Chapter 3 of "Biomedical Applications of Mass Spectrometry, First Supplementary Volume," G.R. Waller and O.C. Dermer, Eds., John Wiley and Sons, Inc., New York, 1980, p. 55.

(5) T.H. Varkony, Y. Shiloach, and D.H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities," J. Chem. Inf. Comp. Sci., 19, 104 (1979).

(6) J.G. Nourse and D.H. Smith, "Nonnumerical Mathematical Methods in the Problem of Stereoisomer Generation," Match, (No. 6), 259 (1979).

(7) N.A.B. Gray, R.E. Carhart, A. Lavanchy, D.H. Smith, T. Varkony, B.G. Buchanan, W.C. White, and L. Creary, "Computerized Mass Spectrum Prediction and Ranking," Anal. Chem., 52 1095 (1980).

(8) A. Lavanchy, T. Varkony, D.H. Smith, N.A.B. Gray, W.C. White, R.E. Carhart, B.G. Buchanan, and C. Djerassi, "Rule-Based Mass Spectrum Prediction and Ranking: Applications to Structure Elucidation of Novel Marine Sterols," Org. Mass Spectrom., 15 355 (1980).

(9) J.G. Nourse, D.H. Smith, and C. Djerassi, "Computer-Assisted Elucidation of Molecular Structure with Stereochemistry," J. Am. Chem. Soc., 102, 6289 (1980).

(10) J.G. Nourse, "Applications of Artificial Intelligence for Chemical Inference. 28. The Configuration Symmetry Group and Its Application to Stereoisomer Generation, Specification, and Enumeration.", J. Amer. Chem. Soc., 101, 1210, (1979).

(11) J.G. Nourse, "Application of the Permutation Group to Stereoisomer Generation for Computer Assisted Structure Elucidation.", in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 19.

(12) J.G. Nourse, "Applications of the Permutation Group in Dynamic Stereochemistry" in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 28.

(13) J.G. Nourse, "Selfinverse and Nonselfinverse Degenerate Isomerizations," J. Am. Chem. Soc., 102, 4883 (1980).

(14) N.A.B. Gray, A. Buchs, D.H. Smith, and C. Djerassi, "Computer-Assisted Structural Interpretation of Mass Spectral Data," Helv. Chim. Acta, 64, 458, (1981).

(15) N.A.B. Gray, C.W. Crandell, J.G. Nourse, D.H. Smith, and C. Djerassi, "Computer-Assisted Interpretation of C-13 Spectral Data," J. Org. Chem., 46 703 (1981).

(16) N.A.B. Gray, J.G. Nourse, C.W. Crandell, D.H. Smith, and C. Djerassi, "Stereochemical Substructure Codes for C-13 Spectral Analysis," Org. Magn. Res., 15, 375 (1981).

(17) R.E. Carhart, D.H. Smith, N.A.B. Gray, J.G. Nourse, and C. Djerassi, "GENOA: A Computer Program for Structure Elucidation Based on Overlapping and Alternative Substructures," j. Org. Chem, 46, 1708 (1981).

(18) D.H. Smith, N.A.B. Gray, J.G. Nourse, and C.W. Crandell, "The DENDRAL PROJECT: Recent Advances in Computer Assisted Structure Elucidation," Anal. Chim. Acta, Computer Techniques and Optimization, 133, 471, (1981).

(19) N. A. B. Gray, "Applications of Artificial Intelligence for Organic Chemistry: Analysis of C-13 Spectra," J. Artificial Intell., in press, (1982).

(20) C. W. Crandell, N. A. B. Gray, D. H. Smith, "Structure Evaluation Using Predicted C-13 Spectra", J. Chem. Inf. Comp. Sci, 22, 48, (1982).

(21) J. Finer-Moore, N. V. Mody, S. W. Pelletier, N. A. B. Gray, C. W. Crandell, D. H. Smith, "Computer-Assisted Carbon-13 Nuclear Magnetic Spectrum Analysis and Structure Prediction for the C-19-Diterpenoid Alkaloids," J. Org. Chem., 46, 3399, (1981).

(22) M. R. Lindley, N. A. B. Gray, D. H. Smith, C. Djerassi, "A Computerized Approach to the Verification of C-13 NMR Spectral Assignments," J. Org. Chem., 47, 1027 (1982).

(23) N. A. B. Gray, "Computer Assisted Analysis of Carbon-13 NMR Spectral Data," Progress in Nuclear Magnetic Resonance Spectroscopy, in press, (1982).

(24) J. G. Nourse, "Specification and Enumeration of Conformations of Chemical Structures for Computer-Assisted Structure Elucidation," J. Chem. Inf. Comp. Sci., 21, 168, (1981).

(25) J. C. Wenger, D. H. Smith, "Deriving Three-Dimensional Representation of Molecular Structure from Connection Tables Augmented with Configuration Designations Using Distance Geometry," J. Chem. Inf. Comp. Sci., 22, 29, (1982).

(26) D.H. Smith, J.G. Nourse, and C.W. Crandell, "Computer Techniques for Representation of Three-Dimensional Substructures and Exploration of Potential Pharmacophores," in Structure Activity Correlation as a Predictive Tool in Toxicology, L. Goldberg, ed., Hemisphere Publishing Corp., New York, in press, (1982).

E. Funding Support

Title:

RESOURCE RELATED RESEARCH: COMPUTERS IN CHEMISTRY (grant)

Principal Investigator:

Carl Djerassi, Professor of Chemistry, Department of Chemistry, Stanford University

Dennis H. Smith (Associate Investigator), Senior Research Associate, Department of Chemistry, Stanford University

Funding Agency:

Biotechnology Resources Program, Division of Research Resources, National Institutes of Health

Grant Identification Number:

RR-00612-13

Total Award and Period:

Total - 5/1/80 - 4/30/83 --------- $641,419

Current Award and Period:

Current - 5/1/82 - 4/30/83 -------- $170,710

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

In the coming period of our research, our computational approaches to structural biochemistry will become much more general and we plan wide dissemination of the programs resulting from our work. These more general approaches to aids for the structural biochemist will yield computer programs with much wider applicability than, for example, the existing CONGEN, GENOA, STEREO and STRCHK programs. We expect that this will create a significant increase in requests for access to our programs, placing heavy emphasis on our relationship with SUMEX to provide this access (see Justification and Requirements for Continued SUMEX Use for additional details).

For these reasons, in our current grant period the SUMEX-AIM resource is identified as the resource to which our research is related. The SUMEX-AIM resource has provided the computational basis for our past program developments and for initial exposure of the scientific community to these programs. The resource is, however, funded completely separately from our own research; we are only one of a nationwide community of users of the SUMEX-AIM facility. Our relationship to SUMEX is one which goes far beyond mere consumption of cycles on the SUMEX machine. It has been the goal of the SUMEX project to provide a computational resource for research in symbolic computational procedures applied to health-related problems. As such research matures, it produces results, among which are computer programs, of potential utility to a broad community of scientists. A second goal of SUMEX has been to promote dissemination of useful results to that community, in part by providing network access to programs running on the SUMEX-AIM facility during their development phases. SUMEX does not, however, have the capacity to support extensive operational use of such programs. It was expected from the beginning that user projects would develop alternative computing resources as operational demands for their programs grew. Such a state has been reached for the CONGEN, GENOA, STEREO and STRCHK programs and future developments in the DENDRAL Project to yield more generally useful programs will simply magnify the problem.

We will, therefore, under our relationship with SUMEX-AIM, participate as before in the SUMEX-AIM community in sharing methods and results with other groups during development of new programs.

A. Scientific Collaboration and Program Dissemination

Scientific Collaborations:

The following is a brief description of collaborative efforts that have been taking place or will soon commence in the use of DENDRAL programs for various aspects of structural analysis.

1) Dr. Raymond Carhart, Lederle Laboratories.

Dr. Carhart (a former member of our group) is engaged in research concerned with computer applications to structure/activity relationships. Program development is done jointly between Lederle and Stanford with free exchange of software. Lederle applications are carried out on their own computer facility.

2) Dr. Janet Finer-Moore, University of Georgia.

Dr. Finer-Moore is engaged in structure analysis of alkaloids in Dr. Peletier's group at Georgia. This research makes extensive use of 13C NMR. Our collaboration involves the development and application of our 13C interpretive and predictive programs in structure elucidation of new compounds based on an extensive set of 13C data available on closely related compounds. Access is via network to our programs at Stanford. We have just published a manuscript as a result of this collaboration. (See Reference 21) (Dr. Finer-Moore has recently moved to the University of California, San Francisco.)

3) Dr. Fred McLafferty, Cornell University.

Dr. McLafferty's research is involved with instrumental and analytical aspects of mass spectrometry. We are working with him on the development and application of an interface between his STIRS system and CONGEN/GENOA for structure determination based on mass spectral data. Part of this collaboration is development of IBM versions of some of our programs. Access is in part to Stanford, shifting primarily to Cornell as development proceeds.

4) Dr. David Cowburn, The Rockefeller University.

Dr. Cowburn's research is in the area of conformational analysis, primarily of peptides. We are working with him on the development and application of our programs for generation of molecular conformations. Dr. Cowburn's works with large ring peptides which represent a significant challenge for a conformation generator. His participation will help assure an eventual program of practical use rather than just theoretical interest. Collaboration will be via network access to our programs at Stanford.

5) Dr. D.C. Rohrer, Medical Foundation of Buffalo.
Research Laboratories, Buffalo, New York.

We have initiated a collaboration with Dr. Rohrer on the problem of finding the common 3-dimensional substructural features of a set of chemical structures. The use of such a program would be to postulate substructural features which are responsible for similar biological or spectral properties. The initial approach is similar to that used successfully to find the greatest common subgraph of a set of constitutional structures. Collaboration will be via network access to Stanford.

6) Dr. Peter Regan - Shell Biosciences Laboratory, Sittingbourne, England.

Dr. Regan has made extensive use of our programs, especially the 13C nmr programs. He has given presentations on the use of CONGEN and has applied the program to the structure determination of a new acidic amino acid, 2,4-methanoglutamic acid, and other compounds from plant seeds. This work was done in collaboration with Prof. Jon Clardy at Cornell.

7) Dr. Margaret Wise - Smith Kline and French Laboratories.

We are collaborating with Dr. Wise on the development of programs for design of new structures. In particular, we are interested in designing GABA (gamma-amino butyric acid) analogues while developing our programs for this purpose (structure design).

8) Dr. Douglas Dorman - Eli Lilly.

Dr. Dorman has been one of our best users. He has given several presentations on the use of CONGEN and other programs. He has used the program as an aid in solving a number of structures including some beta-lactam antibiotic derivatives. He was a trial user of the 13C nmr programs given both his extensive knowledge of nmr and use of our programs. His evaluations of these programs has proven invaluable in their improvement and continued development.

9) Dr. J.N. Shoolery, Varian Associates, Palo Alto.

We are collaborating with Dr. Shoolery and others at Varian to obtain high quality C-13 spectra of several marine sterols available only in very small quantities. This is being done as part of our ongoing project to develop programs which are capable of spectral interpretation and prediction. The Varian people access our programs directly or via network.

Program Dissemination:

We have provided access to our programs to a community of collaborators via 1) distribution of the CONGEN program to other laboratories, and 2) guest or individual accounts on the SUMEX computer facility here at Stanford. These methods to promote the dissemination and use of our programs are elaborated below, followed by a brief description of some of our collaborations.

a) Program Export

The past two years we have distributed CONGEN to a number of laboratories owning computers on which the exportable version can now execute. These currently include DEC PDP-10 and -20 systems operating under the TENEX, TOPS-10 and TOPS-20 operating systems, and more recently, the beginnings of a version for IBM systems. The following persons are currently running CONGEN on their own laboratory computers:

Dr. Larry Anderson - University of Utah
(work described in section on collaborations)

Dr. Hartmut Braun - Organische-Chemisches Institut der Universitat Zurich, Switzerland
A former member of Prof. Wipke's group at UC Santa Cruz. He has only recently installed the program at ETH, Zurich.

Dr. Raymond Carhart - Lederle Laboratories
(work described in section on collaborations)

Dr. Peter Regan - Shell Biosciences Laboratory, England
Dr. Carrington has used the program both as a guest user and recently in export. He has given presentations on the use of CONGEN and has applied the program to the structure determination of a new acidic amino acid, 2,4-methanoglutamic acid, and other compounds from plant seeds. This work was done in collaboration with Prof. Jon Clardy at Cornell who is also a guest user.

Dr. Robert Carter - University of Lund, Sweden
Dr. Carter obtained a version of the program for use of several groups at Universities in Sweden.

Dr. Daniel Chodosh - Smith, Kline & French Laboratories
He has installed CONGEN and written an extensive users' manual for the use of SKF chemists.

Dr. Henry Dayringer - Monsanto Agricultural Products Co.
He and Dr. Schwenzer (now at Gulf) were responsible for obtaining and installing CONGEN. Primary use is as an aid to structure elucidation of photoproducts and metabolites of agricultural chemicals.

Dr. Douglas Dorman - Lilly Research Labs
Dr. Dorman has been one of our best users. He attended our 1978 workshop and has given several presentations on the use of CONGEN. He has used the program as an aid in solving a number of structures including some beta-lactam antibiotic derivatives.

Dr. John Trent - Amoco Standard Oil (Indiana)

Dr. Martin Huber - Ciba-Geigy, Switzerland
Dr. Huber is a former member of Prof. Wipke's group at UC Santa Cruz. He has recently received the program and is currently working to interest his coworkers at Ciba in computer assisted structure elucidation.

Dr. Carroll Johnson - Oak Ridge National Laboratory
Dr. Johnson is a long time colleague who spent a year at Stanford in 1976. He is involved with the analytical group at Oak Ridge and is using the program as an

analytical aid and as a model for programs he is developing.

Dr. G. Jones - ICI Pharmaceuticals, England
He has installed CONGEN and is currently evaluating its utility for use by analytical chemists at ICI.

Dr. Fred W. McLafferty - Cornell University
(work summarized under collaborations)

Dr. Peter W. Milne - CSIRO Division of Computing Research, Australia
He contacted us through his association with the Heuristic Programming Project at Stanford. He has acted as the Australian contact for distribution of CONGEN in that country.

Dr. James Morrison - Latrobe University, Australia
(see Milne, above)

Dr. David Pensak - E.I. duPont de Nemours and Company
(see EXODENDRAL account DUPONT, and workshop)

Dr. Joseph SanFilippo - Rutgers University
Dr. SanFillippo is using CONGEN in conjunction with his work on superoxide chemistry and in the evaluation of mass spectral data for environmental samples.

Dr. William Sieber - Sandoz, Ltd., Switzerland
He has installed CONGEN for use by structural chemists at Sandoz. Currently they are evaluating its utility.

Dr. M.D. Sutherland - University of Queensland, Australia
(see Milne, above)

Dr. R.O. Watts - Australian National University
(see Milne, above)

b) EXODENDRAL Account

We reserve a special account on SUMEX for persons interested in access to our programs. Initially, this account was used for anyone desiring access, independent of expected level of use or eventual interest. As the SUMEX system became more heavily loaded a mechanism for guest access was provided and at that point we began to differentiate our users by level of interest. For those desiring merely to try programs we provide guest access (see page 111). If there is interest in continuing collaboration, EXODENDRAL status is given, which provides access to more system facilities and good file management capabilities. The persons who have been active under EXODENDRAL status this year are the following (with the account name followed by the contact person and association):

<AMOCO>
Dr. John Trent - Standard Oil (Indiana)
Amoco Research Center
A workshop participant with several colleagues. They are using our programs in their analytical division.

<BRAEKMAN>
Dr. Jean-Claude Braekman - Universite Libre de Bruxelles, Belgium
He is a former post doctoral fellow in our group, and accesses CONGEN from Belgium for natural products structure elucidation.

<BRAUN>
Dr. Hartmut Braun - Organische-Chemisches Institut der Universitat Zurich, Switzerland
(see section on export)

<CARRINGTON>
Dr. Peter Regan - Shell Biosciences Laboratory, England
(see section on export)

<COWBURN>
Dr. David Cowburn - The Rockefeller University
(see section on collaborations)

<DORMAN>
Dr. Douglas Dorman - Lilly Research Laboratories
(see section on export)

<DREIDING>
Dr. Andre Dreiding - Organische-Chemisches Institut der Universitat Zurich, Switzerland
He has used CONGEN and STEREO extensively in structural studies. He has also worked closely with Braun (see section on export under Braun).

<DUPONT>
Dr. Earl Abrahamson - E.I. duPont de Nemours and Company
Dr. Abrahamson and 4 colleagues attended our 1980 workshop. They are attempting to integrate our program into their overall computer software system which includes a wide variety of programs for applications to chemical problems.

<FINER-MOORE>
Dr. Janet Finer-Moore - University of Georgia
(see section on collaborations)

<GASH>
Dr. Kenneth Gash - California State College at Dominguez Hills

<HELLER>
Dr. Steven Heller - Environmental Protection Agency
We are continuing our work with the NIH/EPA Chemical Information System, through Heller, to attempt to find mechanisms for making CONGEN accessible through that system.

<HUBER>
Dr. Martin Huber - Ciba-Geigy, Switzerland
(see section on export)

<MILNE>
Dr. Peter W. Milne - CSIRO Division of Computing Research, Australia
(see section on export)

<MONSANTO>
Dr. Henry Dayringer - Monsanto Company
(see section on export)

<MWOOD>
Dr. Mark Wood - Rutgers University

<RCARHART>
Dr. Raymond Carhart - Lederle Laboratories
(see section on collaborations)

<ROHRER>
Dr. Douglas C. Rohrer - Medical Foundation of Buffalo
(see section on collaborations)

<ROUSSEL>
Dr. Jean Mathieu - Roussel UCLAF
(see section on guest access under Delaroff)

<SIEBER>
Dr. William Sieber - Sandoz Ltd., Switzerland
(see section on export)

<SMITHKLINE>
Dr. Margaret Wise - Smith Kline and French Laboratories
(see section on collaborations)

<VARIAN>
Dr. James Shoolery - Varian Associates
(see section on collaborations)

c) GUEST Access

We have provided GUEST access to our programs for those persons desiring occasional access to study a structural problem and for those who

wish a "hands-on" introduction to the programs. Persons who have received information about this method of access and have actually logged in during the past year:

Dr. Robert Adamski - Alcon Labs

Dr. A. Bothner-by - Carnegie Mellon University
Dr. Bothner-by has requested access to aid others in the Chemistry Department with structure elucidation work.

Dr. Reimar Bruening - Institut fur Pharmazeutische Arzneimittellehre der Universitat, West Germany
Dr. Bruening has used the program to aid in his solution of the structure of the alkaloid Cassine. He was a participant in our 1978 workshop and has maintained interest since then. He has given at least one presentation in Germany on our programs.

Dr. William Brugger - International Flavors and Fragrances
Dr. Brugger is interested in eventually obtaining CONGEN for use at IFF in natural products structure elucidation.

Dr. Robert Carter - University of Lund, Sweden
(see section on export)

Dr. Francois Choplin - Institut Le Bel, France

Dr. Jon Clardy - Cornell University
He has used CONGEN on occasion to determine the potential structural variety for an unknown prior to obtaining the X-ray crystal structure.

Dr. Mike Crocco - American Hoechst Corp.

Dr. Dan Dolata - University of California at Santa Cruz
He is one of our contacts with Prof. Wipke's group at UC Santa Cruz.

Dr. Bruno Frei - Laboratorium f. Organische Chemie, Switzerland

Dr. John Gordon - Kent State University
Dr. Gordon has been using CONGEN while working at Chemical Abstracts in Columbus, Ohio. He has been using CONGEN to investigate general issues of structure representation.

Dr. Richard Hogue - University of California at Santa Cruz
He is another contact with Prof. Wipke's group.

Dr. Suba Neir - Washington University, St. Louis
Dr. Neir used the program to aid in determination of the structure of a mutagen.

Dr. A. Neszmelyi - Central Research Institute for Chemistry of the Hungarian Academy of Sciences

Ms. Connie Oshirio - Lawrence Berkeley Labs

Dr. Philip Pfeffer - USDA (Philadelphia)

Dr. Ned Phillips - University of Florida

Dr. J.D. Roberts - California Institute of Technology

Dr. Francis Schmitz - University of Oklahoma

Dr. Michael Zippel - Institut fur Biochemie Zentrale Arbeitsgruppe Spectroskopie, Germany
Dr. Zippel used CONGEN to investigate the possible connection with their spectral search system.

d) Industrial Affiliates Program

The high level of interest shown by industrial research laboratories in our programs has always presented us with delicate questions about access to SUMEX-AIM. In the past we have granted access for trials of our programs under the conditions that access is necessarily limited and that the recording mechanisms of our programs be used to ensure that all such trial use be in the public domain. As of April, 1980, we began solicitation of interested industrial organizations to participate in a DENDRAL Project Industrial Affiliates Program. As of May 1, 1982, we have seven members. We intend to use this program as a means by which we can offer collaborations with our on-going research to industrial organizations separate from SUMEX-AIM. Although EXODENDRAL accounts to such organizations are used to facilitate communication and sharing of new programs and concepts of interest with the community as a whole, all significant and certainly all proprietary use of our programs will be carried out on their own computational facilities.

e) Program License

During the past year we have completed a license agreement with Molecular Design, Ltd. of Hayward, California for exclusive rights to DENDRAL Project programs. We see this mechanism as serving the function of technology transfer in a very realistic way. We do not, as a research project, have the charter or the resources to do what is essentially final engineering of a program and integration of the program into an existing, larger system. Such "value added" effort is crucial to broad acceptance of a computer-based method. In addition, Molecular Design will take on the burden of maintenance, documentation and training, freeing our personnel to pursue our research objectives and to bring experimental programs to the level of performance where they, too, can be disseminated by licenses.

B. Interactions with Other SUMEX-AIM Projects

We routinely collaborate with other projects on SUMEX most closely related to our own research. In particular, these collaborations have taken place with the CRYSALIS project, MOLGEN, SECS and have begun with Dr. Carroll Johnson at Oak Ridge.

CRYSALIS is concerned with new approaches to the interpretation of X-ray crystallographic data. X-ray crystallography is another approach to molecular structure elucidation. One of our long-term interests is exploring ways in which CONGEN or GENOA generated structures might be used to guide the search of electron density maps. We are also communicating with Prof. Jon Clardy at Cornell on this problem. It is hoped that having narrowed down the structural possibilities for an unknown using physical and chemical data, the few remaining candidates can be used to guide interpretation of such maps.

Most of the structural problems investigated by MOLGEN involve much larger molecules than the size normally investigated in DENDRAL research. Thus, structural representations involving higher levels of abstraction are of utility in MOLGEN, making our structure manipulation tasks quite different. However, many of the ways in which MOLGEN manipulates its structural representations drew on past experience in DENDRAL in developing algorithms to perform these manipulations.

We collaborate frequently with the SECS project in a number of ways. Although our research efforts are in one sense directed toward opposite ends of work on chemical structures, SECS being devoted to synthesis, DENDRAL being devoted to analysis, the underlying problems of structural manipulation share many common aspects. We have exchanged software where possible, particularly in the area of chemical structure display. We have held several discussions in joint group meetings and at several symposia including the AIM Workshops on common problems, including substructure searching, canonical representations and representation and manipulation of stereochemistry. Persons visiting one laboratory often take the opportunity to visit the other. For example, recent visitors to both laboratories have included Prof. Andre Dreiding, Zurich, Dr. Martin Huber, Basel, and Prof. Robert Carter, Lund.

Dr. Carroll Johnson has collaborated on the CRYSALIS project in the past. More recently he has taken an interest in the use of knowledge-based programs for certain problems in spectral data interpretation. For this reason he is exploring the AGE and EMYCIN systems as frameworks for his program structure, and is involved in discussions with DENDRAL to see where common areas of data interpretation can be identified so that he can draw on our experience and programs. This effort is just beginning at this time; we plan to meet early in May at Stanford to continue discussions.

C. Critique of Resource Management

The SUMEX-AIM environment, including hardware, system software and staff, has proven absolutely ideal for the development and dissemination of

DENDRAL programs. The virtual memory operating system has greatly facilitated development of large programs. The emphasis on time-sharing and interactive programs has been essential to us in our development of interactive programs. Our experience with other computer facilities has only emphasized the importance of the SUMEX environment for real-world applications of our programs. To run CONGEN, for example, in a batch computing environment would make no sense whatever because the program (and our other, related programs) is successful in large part because an investigator can closely monitor and control the program as it works toward solution. We have no complaints whatsoever about the computing environment.

We do have, however, significant problems with SUMEX-AIM capacity, both in available computer cycles and on-line file storage. In a sense DENDRAL suffers from its success. The rapid progress made during the last grant period and now continuing into the next period has led to development of many new programs as adjuncts to CONGEN and GENOA and at the same time has inspired many persons in the scientific community to request some form of access to our programs. The net result is that it is often very difficult to carry on at the same time development and collaborations involving applications of our programs to structural problems due to high load average on the system.

The current overcrowding we see on SUMEX creates two major problems for us in the conduct of our research. First, it diminishes productivity as many people compete for the resource; the "time-sharing syndrome" leads to idle, wasted time at the terminal waiting for trivial computations to be completed. Second, the slow response time of the system is an aggravation to an outside investigator who is anxiously trying to solve a structural problem. At some point even the most interested persons will give up, log off the computer and resort to manual methods where possible.

We have taken many steps within our project to try to work around heavy use periods on SUMEX. Our group works a staggered schedule, both in terms of the actual hours worked each day and in terms of what days each week are worked. This results in some problems in intra-group communication, but fortunately the message and other communication systems of SUMEX help alleviate that situation. We try to run all demonstrations on the DEC-2020 to help ease the burden on the dual KI-10 system. We encourage our collaborators to avoid prime-time use of the system when possible. We have used our new Megatek Graphics terminal exclusively on the 2020. This terminal makes modest demand on the host computer and it was decided to use it only on the somewhat less used 2020 rather than the dual-KI system

For these reasons, we strongly support the planned augmentation of the SUMEX-AIM hardware. Any part of our computations which can be shifted to another machine will not only facilitate export of our software but will ease the load on the DEC-10s and make it easier to continue our research. Both will serve to make SUMEX more responsive and our productivity higher.

III. RESEARCH PLANS

A. Project Goals and Plans

Current research efforts were described in highlight form in the first section, Summary of Research Program. In this section we discuss in outline form the major goals of our current grant period (5/1/80 - 4/30/83), with an indication of the progress made to date.

Our goals include the following:

1) Develop SASES (Semi-Automated Structure Elucidation System) as a general system for computer aided structural analysis, utilizing stereochemical structural representations as the fundamental structural description. SASES will represent a computer-based "laboratory" for detailed exploration of structural questions on the computer. It will have as key components the following:

A) Capabilities for interpretation of spectral data which, together with inferences from chemical or other data, would be used for determination of (possibly overlapping) substructures. We have made considerable progress in the areas of mass spectrometry (see References 3, 14) and C-13 NMR spectroscopy (see References 15, 18);

B) The GENOA (structure Generation with Overlapping Atoms) program which will have the capability of exhaustive generation of (topological and stereochemical) structural candidates and include as an essential component the existing CONGEN program. We have developed Version I of GENOA for use by our collaborators (see Reference 17);

C) Capabilities for prediction of spectral (and biological) properties to rank-order candidates on the basis of agreement between predicted and observed properties. Again, we have made considerable progress in mass (see References 3, 7, 8) and C-13 NMR (see References 15, 16, 18) spectroscopy;

2) Develop automated approaches to both interpretation and prediction of spectroscopic data, including but not limited to the following spectroscopic techniques:

A) carbon-13 magnetic resonance (13CMR) (see References 15, 16, 18);

B) proton magnetic resonance (1HMR);

C) mass spectrometry (MS) (see References 3, 7, 8);

The interpretive procedures will yield substructural information, including stereochemical features, which can be used to construct structural candidates using GENOA. We have illustrated this method in

recent publications (see References 14, 18, 21). The predictive procedures will be designed to provide approximate but rapid predictions of expected spectroscopic behavior of large numbers of structural candidates, including various conformers of particular structures. Such procedures can be used to rank-order candidates and/or conformers. The predictive procedures will also be designed to provide more detailed predictions of structure/property relationships for known or candidate structures in specific biological applications. These procedures have been illustrated in recent publications (see References 3, 7, 8, 15, 18, 20).

3) Develop a constrained generator of stereoisomers, (see Reference 9) including:

A) design and implement a complete and irredundant generator of possible conformations for a given known, or a candidate for an unknown, structure;

B) provide constraints for the conformation generator so that proposed structures for a known or unknown compound possess only those features allowed by: i) intrinsic structural features such as ring closure and dynamics of the chemical structure; and ii) data sensitive to molecular conformations (e.g., MCD, NMR);

C) integrate the stereochemical developments with the GENOA program as a final, comprehensive solution to the structure generation problem and allow for interface of the program with other methods dependent on atomic coordinates.

4) Promote applications of these new techniques to structural problems of a community of collaborators, including improved methods for structure elucidation and potential new biomedical applications, through resource sharing involving the following methods of access to our facilities and personnel;

A) nationwide computer network access, via the SUMEX-AIM computer resource;

B) exportable versions of programs to specific sites;

C) workshops at Stanford to provide collaborators with access to existing and new developments in computer-assisted structure elucidation in an environment where complex questions of utility and application can be answered directly by our own scientific staff;

D) interface to a commercially available graphics terminal for structural input and output, at as low a cost as possible, so that chemists can draw or visualize structures more simply and intuitively than with our current, teletype-oriented interfaces.

B. Justification and Requirements for Continued SUMEX Use

In previous sections we discussed the relationship between the DENDRAL Project and SUMEX-AIM, methods for using SUMEX-AIM for dissemination of our programs to a broad community of structural chemists and biochemists and a critique of resource management. In this section we wish to emphasize certain factors which were not discussed earlier and to show how our future directions and interests are closely related to the proposed continuation and augmentation of the SUMEX-AIM resource.

As resource-related research, DENDRAL is intimately tied to the SUMEX resource. Our involvement with SUMEX goes far beyond simple use of the facility. We use SUMEX as the focal point for a number of collaborative efforts, for export of our software and for the communication facilities essential to maintaining close contact with remote research groups working with us. SUMEX provides computational facilities for our workshops, where we bring outside investigators to Stanford to use new programs applied to real structural problems. We have already discussed in our critique the difficulties we have, in view of heavy SUMEX load, of maintaining both our research effort and the resource-sharing aspects of our project. To help eas these burdens we are making extensive use of the SUMEX 2020 system via one direct line and via the ETHERNET link from SUMEX to the 2020. Much of our work in graphics software, for example, is carried out on the 2020.

C. Needs and Plans for Other Computing Resources

For several years now we have directed some attention toward alternative computing resources which could be used to support all "production" use of our programs, i.e., all applications designed to use the programs to solve real problems. Although this would have the severe disadvantage of separating our research effort from many of the applications, it has been our hope that emerging technology in networking would enable us to keep in reasonably close contact with another resource. Two resources have emerged as candidates for systems where our programs can be accessed and used in problem-solving. Unfortunately, neither has so far proven feasible for several reasons (mentioned below). At this time we cannot determine if the problems will be resolved. Until such time, we will remain completely dependent on SUMEX for <u>all</u> our computational needs.

One alternative resource is the NIH/EPA Chemical Information System. For more than three years we have been working with them to obtain sufficient contract money to provide a version of CONGEN integrated into that system. The concept and the funds were approved but a contract has never been issued due to administrative problems at the EPA. Although there have been some developments recently, we still have no firm idea on when such a contract will be issued. If this effort is successful, then we can encourage persons who desire access to our programs to consider using the NIH/EPA system.

A second alternative is the National Resource for Computation in Chemistry (NRCC). This Resource has recently had its funding terminated.

We are now pursuing an alternative discussed previously, that of arranging license agreements with private industry for dissemination of our software. This will likely be the focus of our future efforts to disseminate programs to those researchers who merely wish to use them rather than work together with us in collaborative arrangements to develop more powerful programs.

D. Recommendations for Future Resource and Community Development

We have discussed previously our recommendation for the hardware augmentation, particularly with regards to purchase of small machines to facilitate future export. We also have increasing need for more file storage on-line. This is a result of building large data bases as part of our research in spectral interpretation. For the time being we are working with experimental programs and small data bases. As time progresses, however, these data bases will grow rapidly as our group and a number of our collaborators add additional structures and associated spectral data.

Another capability which is of increasing importance to our own work is access to low-cost graphics systems. Our programs will develop increasing dependence on graphics for visualization of three-dimensional molecular structures. Scientists desiring access to our programs will need a graphics terminal for optimum use of our systems. Currently available vector displays are simply too expensive for the average investigator. The emerging technology of low-cost raster display systems offers a more promising possibility. However, no currently available machine has the required capabilities for under $10,000, and this is an area where machines like the Alto hold more promise. SUMEX could perhaps initiate an effort to obtain a system which has the hardware necessary for frame-based display. Such a system allows rotation of three-dimensional objects in a way which permits visualization of the actual shape of the object.

II.A.1.4 EXPEX Project

EXPEX - Expert Explanation Project

Edward H. Shortliffe, M.D., Ph.D.
Departments of Medicine and Computer Science
Stanford University

Michael R. Genesereth, Ph.D.
Computer Science Department
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

EXPEX is not a single project but a combination of efforts that are directed at the development of powerful representational schemes to facilitate knowledge acquisition and explanation. The work includes not only the study of fundamental representational formalisms but also the encoding of various types of knowledge, such as causal information and user models.

We believe that the productivity of basic computer science research tends to be heightened by experiments that deal with significant real world problem domains. Challenges drawn from chemistry, medicine, and molecular biology have introduced additional complexity to expert systems work at Stanford, but have simultaneously forced system developers to respond to pragmatic constraints and user demands that have had a significant impact on the basic AI techniques selected or developed. Thus, we believe that creative investigation into symbolic reasoning techniques is facilitated by working in real world settings where the application forces us to avoid oversimplification. Much of our research effort therefore deals with medical domains (viz., endocrinology and renal pathophysiology) and is being undertaken on SUMEX. Those aspects of the research that deal with nonmedical topics are using other computing resources at Stanford. Our report here will only describe EXPEX, the research on expert medical explanation.

B. Medical Relevance and Collaboration

Our interest in explanation derives from the insights we gained in developing explanatory capabilities for the MYCIN system. In the case of MYCIN and its descendents, we have been able to generate intelligible explanations by taking advantage of its rule-based representation scheme. Rules can be translated into English for display to a user, and their interactions can also be explicitly demonstrated. By adding mechanisms for understanding questions expressed in simple English, we were able to create an interactive system that allowed physicians to convince themselves that they agreed with the basis for the program's recommendations. The

limitations of the explanations generated in this way have become increasingly obvious, however, and have led to improved characterization of the kinds of explanation capabilities that must be developed if clinical consultation systems are to be accepted by physicians.

With this motivation in mind, we are involved in a series of research projects that address medical knowledge representation and explanation. The researchers involved meet regularly to guarantee that each benefits from the insights and progress of the others. We have been fortunate to enlist the collaboration of two additional medical faculty members, Dr. Larry Crapo (endocrinologist), who is helping us build an endocrinology knowledge base, and Dr. Roy Maffly (nephrologist), who is assisting in the development and evaluation of the renal failure work. In the area of endocrinology, the pathophysiology of calcium disorders is the focussed area we are studying because the relationships are well-understood and there are some challenging problems of feedback homeostasis that will need to be represented. The individual projects include the following:

(1) Mr. Greg Cooper's NESTOR program is building on the knowledge base developed for INTERNIST/CADUCEUS and adding causal and temporal relationships. The program is designed to critique a physician's hypothesis regarding the explanation for a set of patient manifestations from the field of hypercalcemia.

(2) Mr. John Kunz is representing the knowledge of renal pathophysiology, including the quantitative relationships that characterize relationships, to develop a consultation and analysis system that melds mathematics and AI techniques.

(3) Mr. Randy Teach is completing a large formal study of the ways in which clinical experts (internists and endocrinologists) explain their findings and recommendations to other physicians. We hope that the results of his analyses will help define the most important characteristics of the explanation components for future medical expert systems.

(4) Dr. Glenn Rennels is working with GUIDON project members on a collaborative effort to implement an explanation capability for the NEOMYCIN program. Because NEOMYCIN's control structure depends on explicit tasking and strategic information, the explanations offered by NEOMYCIN will be very different from those provided by MYCIN's goal-oriented mechanism for rule invocation. Because GUIDON/NEOMYCIN are described in detail in the section of this report dealing with the MYCIN projects, we will not describe this work further in this section.

C. Highlights of Research Progress

THE NESTOR SYSTEM

We have developed a preliminary version of a system that allows a user to input patient data plus an hypothesis, and then have the system critique that hypothesis in light of the data. The system, an evolving thesis project that is largely the work of Mr. Greg Cooper, is called NESTOR and it currently relies completely on information in the INTERNIST-I (CADUCEUS-I) knowledge base.

The more long term goal is to deepen the knowledge in the area of diseases that cause hypercalcemia. For these diseases causal knowledge will be added. A language will be developed to allow the user to state complex time relationships in his description of the patient. NESTOR will be able to accept causal and time statements in hypotheses and reason with them so as to return a hopefully lucid explanation of what it thinks about the user's current hypothesis.

The motivation behind all this research is the deeply help belief that physicians want active control of the diagnostic process and that they also want and need a system that explains, in a user-tailored way, its evaluation of the physician's hypothesis. There may be times when the user wants to give complete control to NESTOR and just be in a mode of answering questions, but we feel that this should be an option and not a requirement. It is observations such as these that have also accounted for the hypothesis assessment work underway in the ONCOCIN research described in the section of this report dealing with the MYCIN projects.

INTEGRATING MATHEMATICAL MODELS WITH AI METHODS

This research project, largely the work of Mr. John Kunz, integrates AI and simple mathematics to analyze a physiological model. In a selected medical domain, we are building a computer program based on these techniques. It analyzes physiological behavior, diagnoses abnormality, and explains the rationale for its analyses. The program fits data to the model, identifies whether the data are abnormal, and identifies the possible causes and effects of any abnormalities. The physiological model is based on knowledge about anatomy, the behavior of the physiological system, and the mechanism of action of the system. The program analyzes many of the problems discussed in Valtin's text Renal Function.

The specific aims of this project are to:

(1) Develop a vocabulary for a physiological model. The vocabulary should represent the "basic physiology" of a biological system. This vocabulary should be adequate to express the concepts included in an introductory professional-level physiology text.

(2) Develop a reasoning system which can solve problems expressed in the vocabulary.

(3) Demonstrate the basic necessity, appropriateness and limitations of the vocabulary and reasoning procedure. This demonstration must be conducted within some limited problem domain.

In this project, mathematics is used in an AI system to compute quantitative values based on available data. Concurrently, AI enhances the simple mathematical models. AI is used to represent diverse knowledge about the problem, to match a solution technique with problems needing solution, and to interpret results. Results may be quantitative and qualitative. Results are explained in terms of their contribution to diagnostic and therapeutic decisions.

This research project was developed within the domain of renal physiology. The project develops a vocabulary for describing a physiological model. In addition to anatomy (structure), this vocabulary describes processes, substances, parameters and mechanisms of action. In general, values of parameters can be measured. Parameters may be related qualitatively or quantitatively. "Mechanism of action" characterizes behavior in terms of laws of physics, specifically including the conservation of mass and Ohm's law. Relations with an associated mechanism of action are characterized as causal. Lacking the mechanism of action, a relation is characterized as behavioral. The mechanism of action provides a very strong focus of attention for the problem solving process. Problems are analyzed in terms of conservation of mass and a search for factors which enable operation of Ohm's law. In addition, the mechanism of action helps to focus the process of acquiring new knowledge about a problem. The behavior, structure and mechanisms of the problem domain are organized as a physiological model.

Using these techniques, the system poses and answers questions such as the following:

> "Is the patient oliguric; define oliguria; what are its causes, consequences and manifestations; what measurements can be made to confirm the diagnosis?"
>
> According to our collaborating nephrologist, Dr. Roy Maffly,
>
> "Appropriate management of patients with excesses or deficiencies of water and the major cations, sodium and potassium, requires an understanding of the basic physiology of these substances in the body. With such an understanding, therapy ordinarily becomes straightforward and logic replaces guesswork." [Maffly, R. Scientific American, 1981].

Early AI diagnosis systems all lack significant knowledge of physiology. The goal of this project is to test the hypothesis that it is possible to develop a computer-based system which uses physiology and an associated "understanding" of that physiology. The system should suggest causes, consequences, and actions which affect physiologically abnormal situations, within a limited medical domain.

PSYCHOLOGICAL STUDIES OF EXPLANATION

This project is aimed at elucidating the nature of expert explanations as they naturally occur in clinical medicine. As we described in last year's annual report, there is remarkably little formal information available regarding the characteristics of high quality explanations among physicians. Thus we decided that it would be valuable to study the way in which experts actually communicate with one another in hopes that these observations will help us arrive at formal design criteria for acceptable expert systems with explanation capabilities.

Mr. Teach brought to bear his background in educational psychology and experimental design to undertake a formal study of this kind. Three general internists and three endocrinologists participated as subjects. With the assistance of Drs. Crapo (endocrinology) and Shortliffe (general medicine), a set of 12 patient management problems was designed. The six subjects were administered the problems in a role-playing "thinking aloud" setting, each case being introduced by a letter from a presumed referring physician. The subjects obtained the history, physical examination, and laboratory information needed to reach a diagnosis and formulate a plan, and they were then asked to dictate a consultation letter back to the referring physician. The referral letters and assignment of cases were carefully controlled to permit optimal control of the variables that might influence the nature and quality of the explanations in the consultation reports. The subjects were not aware that it was the letters themselves (rather than their problem solving behavior) which was a particular area of interest to us. Thus their consultation reports, and the explanations they contain, have provided a basis for the evaluation of the explanations offered by an expert consultant to a physician requesting advice.

The data collection portion of this study is now complete, and the formal statistical analysis of the data is now well underway. A formal dissertation document will be available within the next several months, and several shorter reports on various aspects of the study are expected over the next year.

D. Publications Since January 1981

Wallis, J.W. and Shortliffe, E.H. Explanatory power for expert systems: studies in the representation of causal relationships for medical consultations. Submitted for publication, December 1981.

E. Funding Support

Grant Title: "The Development of Representation Methods to Facilitate Knowledge Acquisition and Exposition in Expert Systems"
Principal Investigator: Edward H. Shortliffe
Agency: Office of Naval Research
ID Number: NR 049-479
Term: January 1981 to December 1983
Total award: $456,622

Grant Title: "Explanatory Patterns in Clinical Medicine"
Principal Investigators: Randy L. Teach and Edward H. Shortliffe
Agency: National Center for Health Services Research
ID Number: 1 R03 HS 04422
Term: March 1, 1981 through April 30, 1982 (extended through August 30, 1982)
Total award: $19,950.00

II. INTERACTION WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations and Program Dissemination via SUMEX

None of these new programs is yet ready for dissemination, although we hope to have prototypes of both NESTOR and the acute renal failure system within the next several months. Our past experience has shown that SUMEX provides a superb vehicle for demonstrating systems, even at a distance.

B. Sharing and Interaction with Other SUMEX-AIM Projects

Although our EXPEX work is young, we are already benefitting from interactions with other researchers who use the SUMEX-AIM resource. The NESTOR work, for example, has depended on access to the INTERNIST-1 knowledge base and on frequent exchange of messages with the researchers at the University of Pittsburgh. Similarly, our collaboration with the GUIDON research team for the implementation of an explanation capability would not have been possible without the facilitated communication and shared file access available via SUMEX.

C. Critique of Resource Management

Although we have not yet placed significant demands on SUMEX management, our previous experience working with Tom Rindfleisch and his staff would suggest that this new project will receive the same kind of laudatory service for which SUMEX has become known.

III. RESEARCH PLANS

A. Project Goals and Plans

THE NESTOR SYSTEM

In the coming year we plan to complete a more sophisticated version of NESTOR. The program will continue to be a decision support tool for diagnosis, but it will be capable of handling much more complicated cases. We envision four major enhancements to the current prototype: